基于随机森林插值的中亚夏季极端高温变化特征

孟欣宁[1],焦瑞莉[1],刘念[2,3],夏江江[2,3*],严中伟[2,3],于爽[2],娄晓[2],
李昊辰[4,5],王立志[2],陈亮[2], 郑子彦[2],赵娜[6]
（1. 北京信息科技大学,北京 100101；2. 中国科学院大气物理研究所,北京 100029；3. 中国科学院大学,北京 100049；4. 北京邮电大学理学院,北京 100876；5. 北京大学,北京 100871；6. 中国科学院地理科学与资源研究所,北京 100101）

摘要：利用中亚地区 65 个气象站的逐日最高气温数据，结合 ERA-Interim 再分析资料以及经纬度、海拔数据，构建了随机森林插值模型，并验证了其可靠性。基于该模型补全了气象站缺失值，获得完整的站点逐日最高气温数据集 $T_{\mathrm{Station_f}}$，并插值得到中亚 1979—2016 年空间分辨率为 0.75°×0.75°的逐日最高气温格点数据集 $T_{\mathrm{RFIM_G}}$。基于 $T_{\mathrm{RFIM_G}}$ 进一步分析了中亚 1979—2016 年夏季极端高温指数时空变化特征。结果表明：中亚区域平均极端高温指数增速在 0.22—0.30°C·(10a)$^{-1}$，显著增温的区域主要分布在哈萨克斯坦的西部、土库曼斯坦大部、乌兹别克斯坦东南部等地区。基于 $T_{\mathrm{RFIM_G}}$ 得到的夏季极端高温指数增速显著大于基于 $T_{\mathrm{Station_f}}$ 得到的结果，这表明用站点观测数据对该地区夏季极端高温趋势的估计明显偏低。本研究得到的数据集可在一定程度上弥补使用站点观测数据片面刻画中亚地区极端高温变化的缺陷，有助于更确切地引导人们在应对极端天气气候事件时采取相应的减缓和适应措施。

关键词：随机森林插值；机器学习；夏季极端高温；中亚
[1]

极端高温事件的增多可对人体健康、生态系统和社会经济带来负面影响[1-2]。近年来，全球极端高温事件都有增多的趋势[3]。中亚地区（哈萨克斯坦、塔吉克斯坦、吉尔吉斯斯坦、乌兹别克斯坦、土库曼斯坦，图 1）作为“一带一路”倡议的核心区域之一，气温变化剧烈，海拔差异较大，夏季易出现高温天气[4]，具有绿洲-荒漠格局的空间异质性[5]，在全球气候变化中很容易发生快速的水文等地貌变化，是全球对气候变化最敏感的区域之一[6]。研究中亚夏季极端高温变化不仅可以加深对该地区气候变化规律本身的认识，也有望在此基础上提出针对该地区应对极端高温事件的减缓和适应措施，确保“一带一路”倡议的可持续发展。

以往的研究表明，近几十年中亚的平均气温具有稳步上升的趋势[7]，且相对于北半球具有更快的增温趋势[8]，其历史极端高温事件如热浪发生频率、强度和持续时间也在不断增加[9-10]。

这些研究所使用的数据或基于气象观测站数据，或基于格点数据集。但是中亚地区可用

收稿日期：2019-10-09；**修订日期**：2020-03-27
基金项目：中国科学院战略性先导科技专项（A 类）资助(XDA20020201)
作者简介：孟欣宁(1993-)，女，硕士，研究方向为信号处理与网络计算. E-mail: 18801036602@163.com
通讯作者：夏江江. E-mail: xiajj@tea.ac.cn

气象观测站少，且空间分布不均，基于稀疏站点数据的分析结果难以代表区域气候变化；虽然格点气温数据集可代替站点数据分析区域极端天气气候事件的变化特征，但以往研究得到的中亚地区格点气温数据集或者非逐日尺度[11]，或者非逐日最高气温[12-14]，或者时间长度不够[15]，无法用以分析中亚夏季极端高温事件的变化特征。这导致了在过去百年全球变暖的大背景下，没有可用的高时空分辨率数据细致、准确地刻画历史中亚夏季极端高温的变化特征。因此，为了更好地了解中亚夏季极端高温的变化特征，首先需要将现有稀疏的站点最高气温数据插值到格点，得到高时空分辨率的网格化逐日最高气温数据集。

气象学和气候学中最常用的插值技术有：最近邻法、样条、回归和克里金法等[15]。这些传统方法多基于统计方法，即需要主观先验知识的代入，其数据或变量通常是特定的，这可能因未完全理解物理过程，导致得到的插值数据集存在缺陷[16]。

相比于传统插值方法，机器学习技术基于自适应机制，不仅可以从数据中学习，不依靠假设，而且可以捕获数据中未知或难以描述的功能关系，并很容易地处理多种不同的数据源，正逐渐成为常用的插值技术。以往研究表明，机器学习中的随机森林算法在环境变量[16]和逐月气温[17]的空间插值研究中表现出色。

本研究以 ERA-Interim 再分析数据、海拔和经纬度数据等作为输入（特征），65 个中亚站点逐日最高气温作为输出（标签），采用随机森林算法构建插值模型，对中亚逐日最高气温站点缺测数据补全和格点数据空间插值，并分析该地区夏季极端高温的变化特征。

1. 研究数据及处理

1.1 站点观测数据

研究中用到的站点逐日最高气温数据来自 NCDC（ftp://ftp.ncdc.noaa.gov/pub/data /gsod）和 GHCN-D（https://www1.ncdc.no- aa.gov/pub/data/ghcn/daily/）。

首先对这两套数据进行质量控制：如果某一站点某一年 6—8 月（夏季）逐日最高气温数据缺测百分比大于 14%，则定义该站点该年不可用；同时，对单个观测站 1979—2016 年有 10%以上的年份不可用时，舍弃该站点的数据。最终得到 65 个观测站（图 1）1979—2016 年夏季（6—8 月）的逐日最高气温数据，各站点多年缺测情况如图 2 所示。

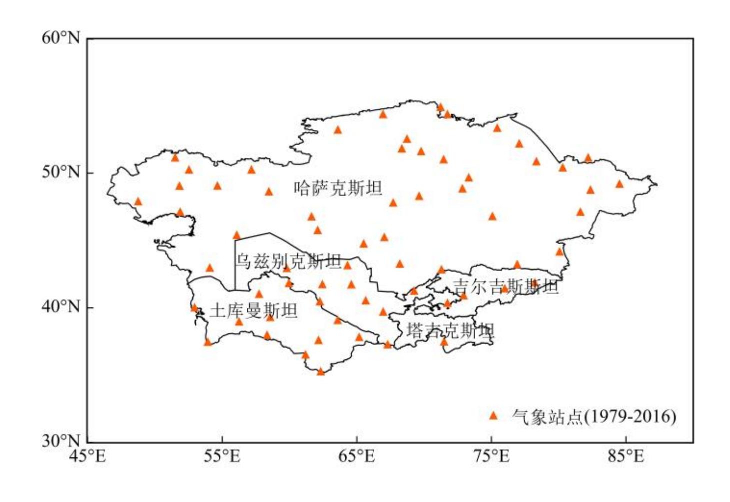


图 1 中亚五国范围及气象观测站点空间分布示意图

Fig.1 Range of Central Asia and the spatial distribution of the meteorological stations

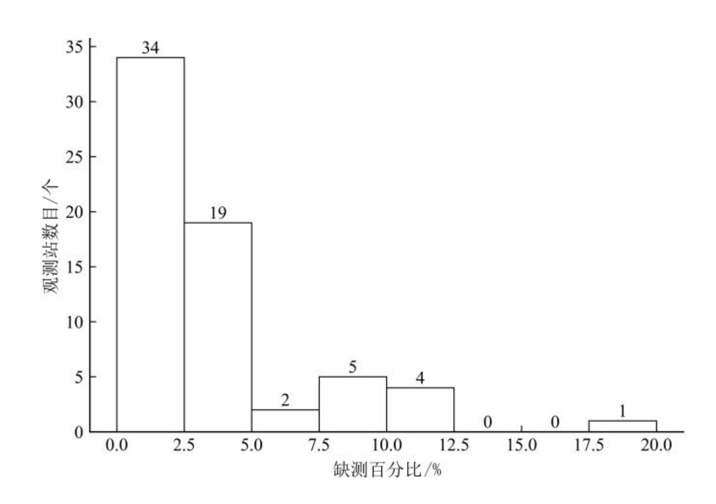


图 2 65 个气象站观测数据缺测情况统计

Fig.2 Statistics on missing observation data of 65 meteorological stations

1.2 再分析数据

再分析数据是在多种来源的观测数据驱动下，利用数值天气预报模式和资料同化技术得到的历史格网气象数据[18]。由于再分析数据具有空间连续性，因此，可以克服站点数据稀少的限制，成为区域气候变化研究的重要数据源[18]。

本研究采用来自欧洲中期天气预报中心的全球大气再分析数据（ERA-Interim），该数据集在全球不同区域都能很好地匹配站点观测数据[5]，因其较高的空间精度和质量[19]而适用于中亚的气候研究。ERA-Interim 空间分辨率为 0.75°×0.75°，共包含 48 种气象要素（https://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/），本文只提取气温最高的时次（12:00）对应的所有气象要素数据，时间段为 1979—2016 年。

表 1 各数据集信息

Tab.1 Dataset information used in this study

名称	数据源	时段	空间分辨率
观测数据	NCDC/GHCN-D	1979—2016	–
ERA-Interim	ECMWF	1979—2016	0.75°×0.75°
经纬度	NCDC/GHCN-D	–	0.75°×0.75°
海拔	NCDC/GHCN-D	–	0.75°×0.75°

2. 研究方法

2.1 插补/插值算法及评估方法

本研究采用随机森林（random forest, RF）对中亚地区逐日最高气温进行“预测”，即进行逐日最高气温数据的站点插补/格点插值。RF 是一种基于决策树的集成学习算法[20-21]。对于回归问题，其最终预测结果是多个决策树预测值的均值。该算法具有计算速度快，鲁棒性高，不易产生过拟合等优点[22]。本研究将通过网格搜索（grid search）方法调节 RF 参数[23]，并使用均方根误差（root mean square error, RMSE）评估模型性能。利用该算法训练出最优模型对站点缺测数据补全和格点数据空间插值。

2.2 研究方案

将 65 个站点的逐日最高气温、经度、纬度、海拔（表 1）、各站点观测气温对应的年、月、日等 7 个要素，以及距离观测站点最近的 ERA-Interim 格点所有 48 个气象要素数据，共同构成模型的训练、验证、测试数据集，用于模型的训练、评估和优化。其中站点逐日最高气温是输出的被解释变量（标签），其余 54 个要素为模型输入的解释变量（特征）。

2.2.1 RFIM 模型的建立与评估

模型构建使用站点逐日最高气温及其对应特征完成（图 3）。将站点观测值及其特征划分为 3 部分：训练集（80%）、验证集（20%）和测试集（2016 年）。通过将训练数据集用于训练模型，可以得到最初的 RFIM（random forest interpolation model）预测模型，但该模型往往拟合效果不佳，还需要以验证集误差为参考，调节模型超参数，如最大特征数、子树数量等。当模型在验证集上的 RMSE 不再明显下降时，表明模型在验证集上的拟合效果达到最佳，然后将测试数据集输入模型以评估其泛化能力：用模型模拟预测 2016 年中亚各站点逐日最高气温，与历史真实观测值对比，求 RMSE 并评估。最后，保存建立好的 RFIM 模型，用于后续站点缺测数据补全和格点数据的空间插值。

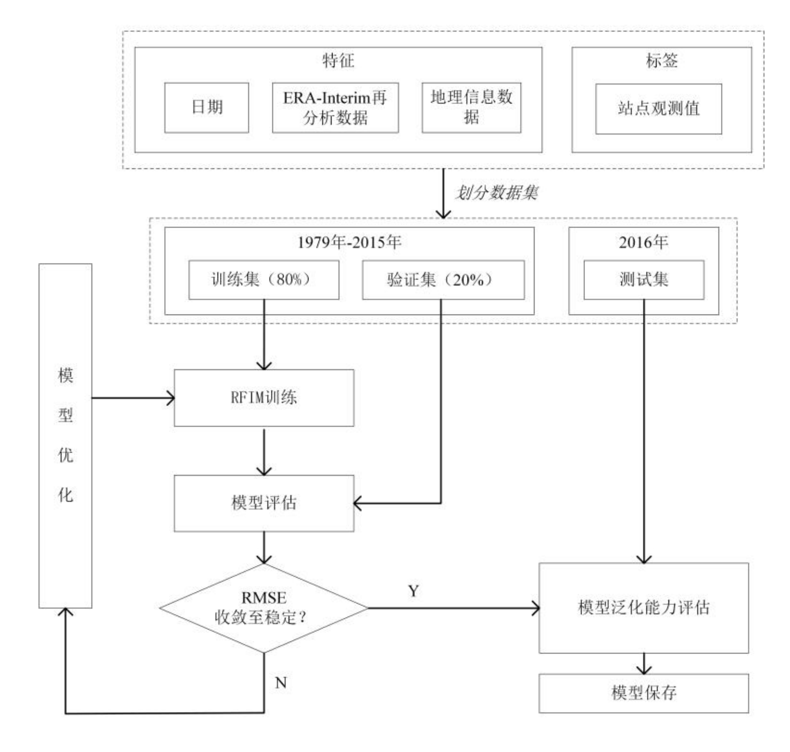


图 3　随机森林插值模型（RFIM）技术路线图

Fig.3 Random Forest Interpolation Model (RFIM) Technology Roadmap.

为后文表述方便，这里定义站点逐日最高气温观测值为 $T_{Station}$，距离观测站点最近的 ERA-Interim 格点气温为 T_{ERA}，RFIM 预测的站点逐日最高气温为 T_{RFIM_S}，RFIM 预测的格点逐日最高气温为 T_{RFIM_G}，补全后的站点逐日最高气温数据集为 $T_{Station_f}$

2.2.2　站点缺测数据补全和格点数据的空间插值

对于站点的缺测数据，取距离缺测站点最近的格点解释变量作为替代输入到 RFIM 模型，预测得到缺测的站点逐日最高气温，即可补全各个站点缺测数据。将中亚各格点的解释变量输入 RFIM 模型，即可插值得到中亚 0.75°×0.75°空间分辨率的格点逐日最高气温数据 T_{RFIM_G}，基于 T_{RFIM_G} 研究中亚地区夏季极端高温变化特征。

3.　研究结果

3.1　RFIM 模型性能评估

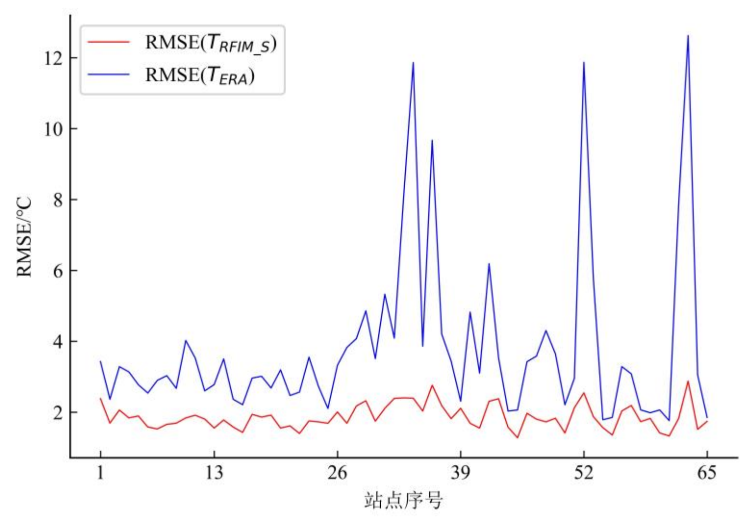


图 4 中亚各站点 2016 年 6—8 月 T_{Station} 分别与 $T_{\mathrm{RFIM_S}}$、T_{ERA} 的 RMSE

Fig.4 RMSE between $\mathrm{T_{RFIM_S}}$ and T_{Station}, between T_{ERA} and T_{Station} for the Summer 2016 in Central Aisa

由图 4 可知，RFIM 模型预测得到站点逐日最高气温（$T_{\mathrm{RFIM_S}}$）与站点观测值（T_{Station}）的 RMSE（65 站平均 RMSE 为 1.87 °C）显著低于 ERA-Interim 格点最高气温（T_{ERA}）与站点观测值（T_{Station}）之间的 RMSE（65 站平均 RMSE 为 3.81 °C）。同时，T_{RFIM_S} 与 T_{Station} 的 RMSE 在各站点之间没有出现较大的波动，表明 RFIM 模型预测结果具有较高的可靠性和稳定性。

为了更直观地验证 RFIM 模型预测的准确性，对 RFIM 预测得到的 2016 年 65 个观测站点的逐日最高气温（T_{RFIM_S}）平均与距离站点最近的 ERA-Interim 格点 2 m 温度（T_{ERA}）进行对比（图 5）。从图 5 可以看出，在区域平均水平上，T_{ERA} 表现出对站点观测气温的低估，这与前人研究结论一致[24]。而模型预测的站点结果 T_{RFIM_S} 和站点观测结果 $T_{Station}$ 基本趋于一致，证明了 RFIM 模型的预测结果更接近站点真实观测值，可用于中亚地区气候变化分析。

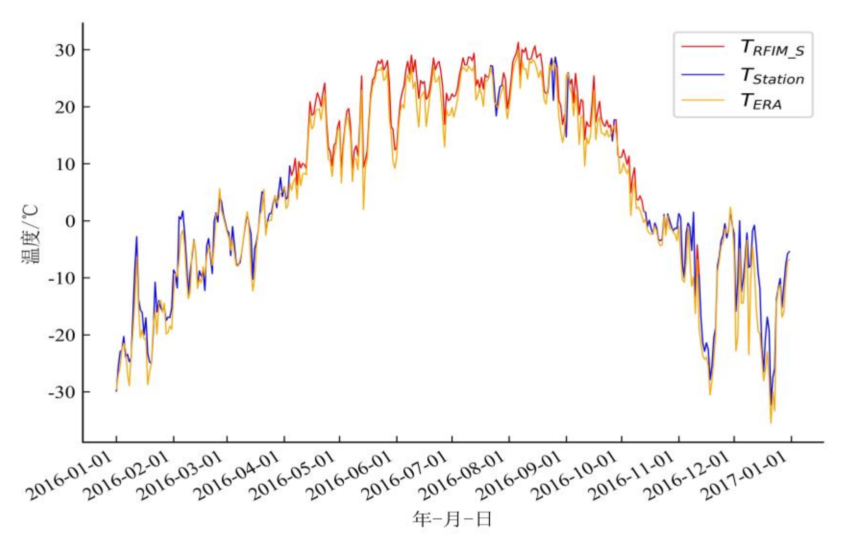


图 5　2016 年逐日最高气温中亚区域平均：$T_{Station}$，T_{RFIM_S}，T_{ERA}

Fig.5 Regional average of daily maximum temperatures in 2016: $T_{Station}$, T_{RFIM_S}, T_{ERA}

3.2　中亚地区夏季极端高温变化特征

利用 RFIM 模型得到 1979—2016 年中亚逐日最高气温 $T_{\mathrm{RFIM_G}}$ 格点资料（2.2.2）。在此基础上对各格点取每年夏季 6—8 月前 *n*（*n* 取 1，5，10，15）个逐日最高气温的平均值，记为指数 TXn。以这 4 个指数作为夏季极端高温强度指数，计算其线性趋势，并采用滑动 *t* 检验[25]（显著性水平 α=0.05）对其进行显著性检验。同时，计算基于 $T_{\mathrm{Station_f}}$ 的 TXn 以作对比。

如图 6 所示，基于 $\mathrm{T_{RFIM_G}}$ 得到的中亚区域平均夏季极端高温强度有增加的趋势，TX1、TX5、TX10 和 TX15 的增速分别为 0.22 ℃ • $(10a)^{-1}$、0.27 ℃ • $(10a)^{-1}$、0.30 ℃ • $(10a)^{-1}$ 和 0.30 ℃ • $(10a)^{-1}$，其值依次增大，这表明更具“平均”意义的夏季极端高温指数增速大于更为“极端”意义的夏季极端高温指数增速，4 个指数的线性趋势均通过显著性检验；相应地，基于 $T_{\mathrm{Station_f}}$ 得到的 TX1、TX5、TX10 和 TX15 的增速分别为 0.02 ℃ •$(10a)^{-1}$、0.12 ℃ •$(10a)^{-1}$、0.16 ℃ • $(10a)^{-1}$ 和 0.19 ℃ • $(10a)^{-1}$，但是只有 TX15 指数的变化趋势通过显著性检验。这也说明，只用站点逐日最高气温数据计算中亚夏季极端高温强度，将会显著低估其增加趋势。

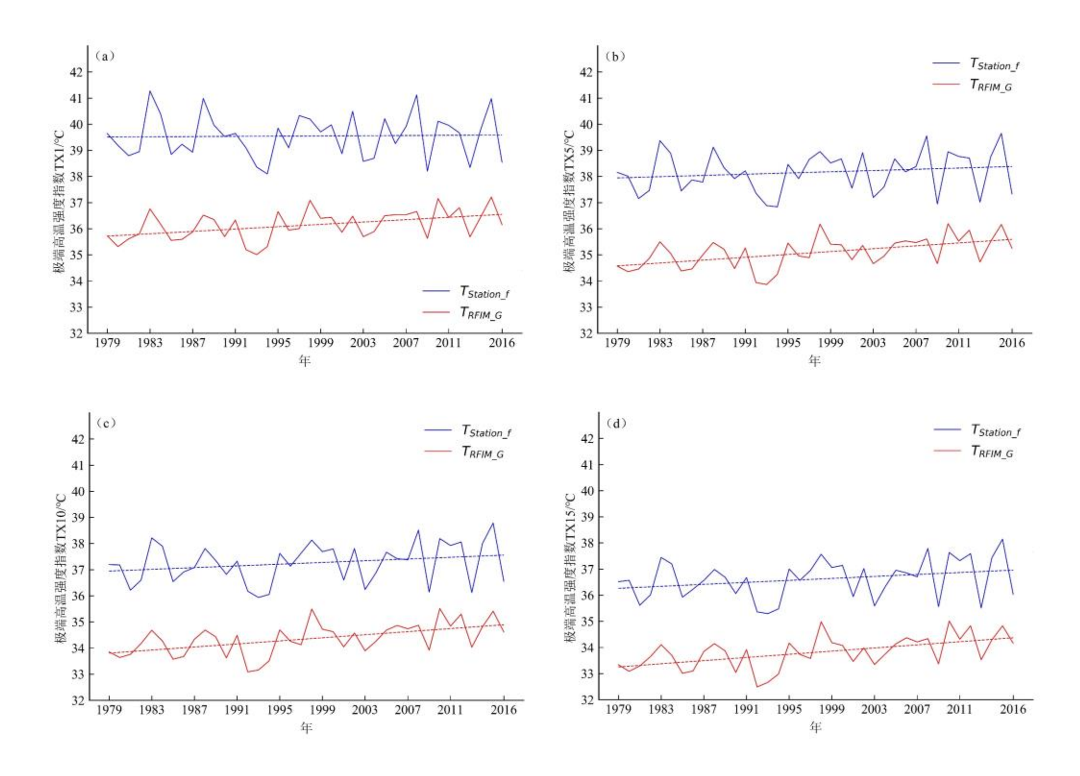


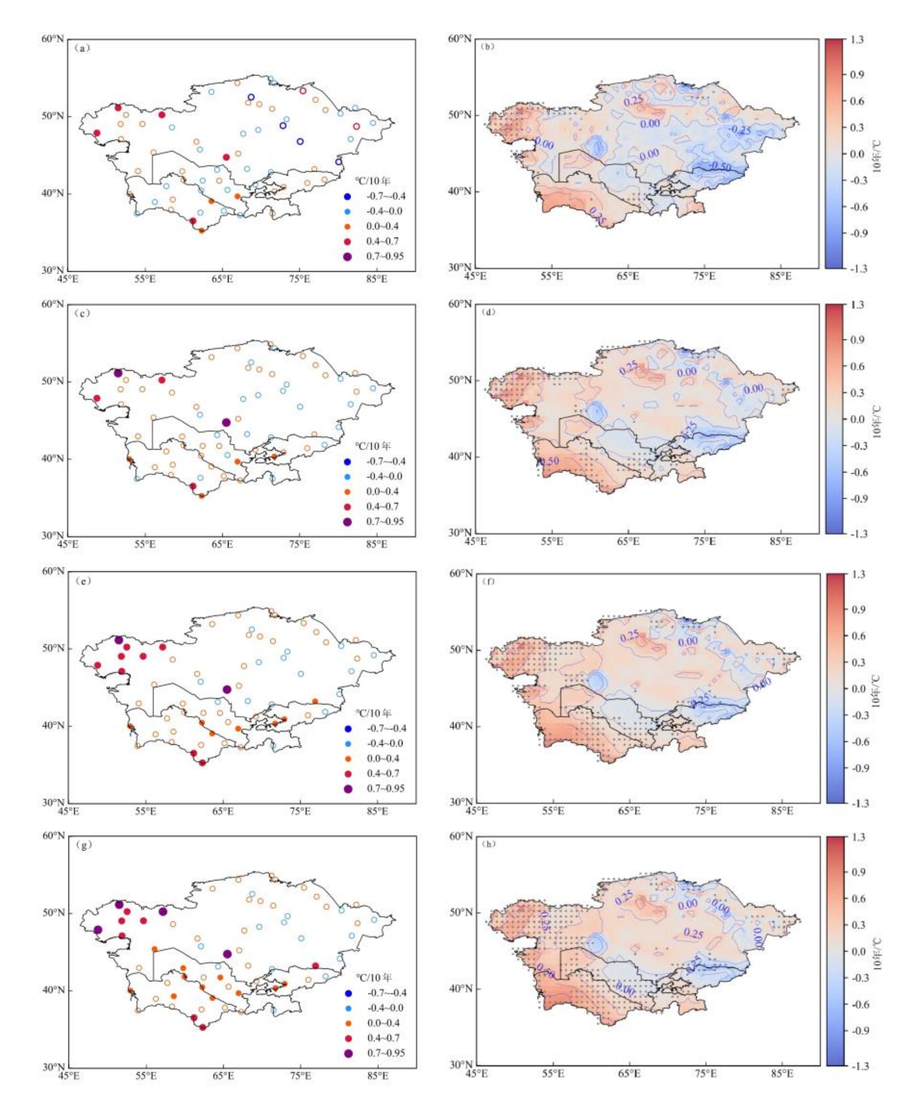
图 6 1979—2016 年区域夏季平均极端高温强度指数变化

Fig.6 The annual time series of extreme high temperature indices averaged over Central Asia in summer

由图 7 可知，基于 T_{RFIM_G} 计算得到的 1979—2016 年中亚大部分地区 4 个夏季极端高温强度变化趋势空间分布比较一致，且同样验证了更具“平均”意义的夏季极端高温指数增速大于更为“极端”意义的夏季极端高温指数增速：显著增温的区域主要分布在哈萨克斯坦的西部、土库曼斯坦大部、乌兹别克斯坦东南部；而塔吉克斯坦夏季极端高温强度增温趋势不显著；吉尔吉斯斯坦大部分地区 4 个夏季极端高温强度指数都为减小趋势，但 TX1、TX5 和 TX10 指数变化未通过显著性检验，部分区域 TX15 指数线性变化通过显著性检验。

以 TX15 的结果为例，上述有显著增温的区域如哈萨克斯坦的西部、土库曼斯坦大部、乌兹别克斯坦东南部区域平均 TX15 的变化趋势分别为 0.6 ℃·$(10a)^{-1}$、0.30 ℃·$(10a)^{-1}$、0.32 ℃·$(10a)^{-1}$。对中亚地区夏季平均气温的变化特征已有的研究中也表明，该地区夏季西部升温比东部快[6],[26]，与本研究结果较为一致。

基于 $T_{\mathrm{Station_f}}$（图 7）计算得到的 4 个极端高温强度指数中分别有 8、9、17、22 个站点通过显著性检验，其变化趋势的空间分布也与基于 $T_{\mathrm{RFIM_G}}$ 结果的空间分布基本一致，再次证实了格点空间插值数据的可靠性。但可以明显看出，基于有限站点数据得到的结果无法细致刻画区域尺度趋势变化的空间分布特征。



注：(a)，(c)，(e)，(g) 为基于 $T_{Station_f}$ 的结果，分别对应 TX1、TX5、TX10、TX15，实心站点代表通过显著性水平 α=0.05 的显著性检验；(b)，(d)，(f)，(h) 为基于 T_{RFIM_G} 的结果，分别对应 TX1、TX5、TX10、TX15，'•'代表通过显著性水平 α=0.05 的显著性检验。

图 7　1979—2016 年中亚夏季极端高温线性趋势的空间分布

Fig.7 Spatial distribution of linear trends of extreme high temperature indices in Central Asia

4. 结论

以往对气温插值和气候变化及其影响的研究大多直接使用站点观测数据进行[26-27]，这样的方式难以细致刻画站点稀疏区域的气候变化情况。虽然有一部分学者采用了网格化的气象数据[5,28-29]，但这些数据要么基于传统插值方法得到，要么直接使用再分析资料开展，而这种基于统计方法的方式容易导致插值数据存在主观性，同时由于再分析资料易受到数值资料、数据同化方法的影响，容易引入虚假的气候变化信号影响分析结果[30]。

本研究利用中亚地区 65 个气象站的逐日最高气温数据，结合 ERA-Interim 再分析资料以及经纬度、海拔数据，构建了随机森林插值模型，得到了中亚地区 1979—2016 年空间分辨率为 0.75°×0.75°的逐日最高气温格点数据集 $T_{\mathrm{RFIM_G}}$。基于此分析了中亚夏季极端高温区

域平均变化趋势，结果表明，该地区夏季极端高温强度增加，其增加速率显著大于基于 $T_{Station_f}$ 得到的结果。这表明用有限的站点观测数据会明显低估该地区夏季极端高温趋势。这将导致人们对中亚历史气候变化认识不足，以至于低估气候变化造成的相关极端天气气候事件（高温、热浪等）所带来的风险，从而导致人们在应对极端天气气候事件时所采取的减缓或适应措施难以达到预期效果。此外，针对本文提出的这种基于随机森林的方法，不仅能够保证插值数据的客观性，而且能够在一定程度上避免再分析资料引入的分析误差，提高插值数据的准确性，即使在数据稀疏区域也可以开展高时空分辨率下的气候变化研究。

参考文献（References）：

[1] Yu Shuang, Xia Jiangjiang, Yan Zhongwei, et al. Loss of work productivity in a warming world: Differences between developed and developing countries[J]. Journal of Cleaner Production,2019,208(6):1219-1225.

[2] Xia Jiangjiang, Tu Kai, Yan Zhongwei, et al. The super-heat wave in eastern China during July-August 2013: a perspective of climate change[J]. International Journal of Climatology,2016, 36(3):1291-1298.

[3] 焦文慧，张勃，黄涛，等. 近 30a 河东地区极端气温时空变化[J]. 干旱区研究, 2019, 36(6): 1466-1477. [Jiao Wenhui, Zhang Bo, Huang Tao, et al. Spatiotemporal change of extreme temperature in the Hedong Region in recent 30 years[J]. Arid Zone Research, 2019, 36(6): 1466-1477.]

[4] 龚子同，陈鸿昭，杨帆，等. 中亚干旱区土壤地球化学和环境[J]. 干旱区研究, 2017, 34(1): 1-9. [Gong Zitong, Chen Hongzhao, Yang Fan, et al. Pedogeochemistry and environment of aridisol regions in Central Asia[J]. Arid Zone Research,2017, 34(1): 1-9.]

[5] 徐婷，邵华，张弛. 近 32 a 中亚地区气温时空格局分析[J]. 干旱区地理, 2015, 38(1): 25-35. [Xu Ting, Shao Hua, Zhang Chi. Temporal pattern analysis of air temperature change in Central Asia during 1980-2011[J]. Arid Land Geography, 2015, 38(1): 25-35]

[6] 沈伟峰，缪启龙，魏铁鑫，等. 中亚地区近130多a气温变化特征[J]. 干旱气象, 2013, 31(1): 32-36. [Shen Weifeng, Miao Qilong,Wei Tiexin, et al. Analysis of temperature variation in recent 130 years in Central Asia[J]. Journal of Arid Meteorology, 2013, 31(1): 32-36.]

[7] Lioubimtseva E, Cole R. Uncertainties of climate change in arid environments of Central Asia[J]. Reviews in Fisheries Science, 2006, 14(1-2):29-49.

[8] IPCC. Climate Change 2013: The Physical Science Basis[M]. Cambridge, UK: Cambridge University Press, 2013: 159-254.

[9] Perkins S E, Alexander L V, Nairn J R. Increasing frequency, intensity and duration of observed global heatwaves and warm spells[J]. Geophysical Research Letters, 2012, 39(20): L20714.

[10] Yu S, Yan Z W, Freychet N, et al. Trends in summer heatwaves in central Asia from 1917 to 2016: Association with large-scale atmospheric circulation patterns[J]. International Journal of Climatology: 2020, 9(1): 115-127.

[11] Alexander L V, Zhang X B, Peterson T C, et al. Global observed changes in daily climate extremes of temperature and precipitation[J]. Journal of Geophysical Research: Atmospheres, 2006, 111(D5): 1042-1063.

[12] Piper S C, Stewart E F. A gridded global data set of daily temperature and precipitation for terrestrial biospheric modeling[J]. Global Biogeochemical Cycles, 1996, 10(4): 757-782.

[13] New M, Lister D, Hulme M, et al. A high resolution data set of surface climate over global land areas[J]. Climate Research, 2002, 21(1): 1-25.

[14] Hijmans R J, Cameron S E, Parra J L, et al. Very high resolution interpolated climate surfaces for global land areas[J]. International Journal of Climatology, 2005, 25(15): 1965-1978.
[15] Kilibarda M, Hengl T, Heuvelink G B M, et al. Spatio-temporal interpolation of daily temperatures for global land areas at 1km resolution[J]. Journal of Geophysical Research: Atmospheres, 2014, 119(5): 2294-2313.
[16] Li J, Heap A D, Potter A, et al. Application of machine learning methods to spatial interpolation of environmental variables[J]. Environmental Modelling & Software, 2011, 26(12): 1647-1659.
[17] Appelhans T, Mwangomo E, Hardy Douglas R, et al. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania[J]. Spatial Statistics, 2015, 14(5): 91-113.
[18] 范彬彬，罗格平，张弛，等. 新疆夏季降水时空分布的适用性评估[J]. 地理研究, 2013, 32(9): 1602-1612. [Fan Binbin, Luo Geping, Zhang Chi, et al. Evaluation of summer precipitation of CFSR, ERA-Interim and MERRA reanalyses in Xinjiang[J]. Geographical Research, 2013, 32(9): 1602-1612.]
[19] Dee D P, Uppala S M, Simmons A J, et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system[J]. Quarterly Journal of the Royal Meteorological Society, 2011, 137(656): 553-597.
[20] 马慧娟,高小红,谷晓天.随机森林方法支持的复杂地形区土地利用/土地覆被分类研究[J]. 地球信息科学学报, 2019, 21(3): 359-371. [Ma Huijuan, Gao Xiaohong, Gu Xiaotian. Random forest classification of landsat 8imagery for the complex terrain area based on the combination of spectral, Topographic and Texture Information[J]. Journal of Geo-information Science, 2019, 21(3): 359-371.]
[21] 王奕森，夏树涛. 集成学习之随机森林算法综述[J].信息通信术，2018, 12(1): 49-55. [Wang Yisen, Xia Shutao. A survey of random forests algorithms[J]. Information and Communications Technologies, 2018, 12(1): 49-55.]
[22] 温小乐，钟奥，胡秀娟. 基于随机森林特征选择的城市绿化乔木树种分类[J]. 地球信息科学学报,2018, 20(12): 1 777-1 786. [Wen Xiaole, Zhong Ao, Hu Xiujuan. The classification of urban greening tree species based on feature selection of random forest[J]. Journal of Geo-information Science, 2018, 20(12): 1777-1786.]
[23] 崔东文，金波. 基于随机森林回归算法的水生态文明综合评价[J]. 水利水电科技进展，2014, 34(5): 56-60+79. [Cui Dongwen, Jin Bo. Comprehensive evaluation of water ecological civilization based on random forests regression algorithm[J]. Advances in Science and Technology of Water Resources, 2014, 34(5): 56-60+79.]
[24] 陈涛，智海，边多. 青藏高原观测地表温度与 ERA-Interim 再分析资料的差异及归因分析[J]. 山地学报, 2019, 37(1): 1-8. [Chen Tao, Zhi Hai, Bian Duo. Investigation on the discrepancy between observed surface temperature and ERA-Interim over the Qinghai-Tibet Plateau andits attribution[J]. Mountain Research, 2019,37(1): 1-8.]
[25] 董光辉，赵柳入，黄海波,等. 假设检验在供应商变更中的应用[J]. 中国药事, 2017, 31(10): 1142-1146. [Dong Guanghui, Zhao Liuru, Huang Haibo, et al. Application of hypothesis test in supplier change[J]. Chinese Pharmaceutical Affairs, 2017, 31(10): 1142-1146.]
[26] 张影，徐建华，陈忠升，等. 中亚地区气温变化的时空特征分析[J]. 干旱区资源与环境，2016, 30 (7): 133-137. [Zhang Ying, Xu Jianhua, Chen Zhongsheng, et al. Spatial and temporal variation of temperature in Central Asia[J]. Journal of Arid Land Resources and Environment, 2016, 30(7): 133-137.]
[27] 沈皓俊，游庆龙，王朋岭，等. 1961—2014 年中国高温热浪变化特征分析[J]. 气象科学, 2018, 38(1): 28-36. [SHEN Haojun, You Qinglong, WANG Pengling, et al. Analysis on heat waves variation features in China during 1961—2014[J]. Journal of the Meteorological Sciences, 2018, 38(1): 28-36.]
[28] 聂羽，韩振宇，韩荣青，等. 中国夏季热浪持续天数的年际变化及环流异常分析[J].气象，2018, 44(2):

294-303. [Nie Yu, Han Zhenyu, Han Rongqing, et al. Interannual variation of heat wave frequency persistence over china and the associated atmospheric circulation anomaly[J]. Meteorological monthly, 2018, 44(2): 294-303.]

[29] 金红梅，颜鹏程，柏庆顺，等.近 70 a 来中亚极端高温事件时空分布[J]. 干旱气象, 2019, 37(4): 550-556. [Jin Hongmei, Yan Pengcheng, Bai Qingshun, et al. Spatial and temporal distribution of extreme high temperature events in Central Asia over the last 70 years[J]. Journal of Arid Meteorology, 2019, 37(04): 550-556.]

[30] 海日古丽·纳麦提，玉素甫江·如素力，玛地尼亚提·地里夏提，等. ERA-Interim 和 GHCN-CAM 再分析气温数据在天山山区的适应性分析[J]. 山地学报, 2019, 37(4): 613-621. [Hairiguli Namaiti, Yusufujiang Rusuli,Madiniyati Dilixiati, et al. Adaptability analysis of ERA-Interim and GHCN-CAM reanalyzed data temperature values in Tianshan Mountains Area, China[J]. Mountain Research, 2019, 37(4): 613-621.]

Change of summer extreme high temperature in Central Asia based on interpolated data by random forest

MENG Xin-ning[1], JIAO Rui-li[1], LIU Nian[2,3], XIA Jiang-jiang[2,3*], YAN Zhong-wei[2,3], YU Shuang[2,3], LOU Xiao[2], LI Hao-chen[4,5], WANG Li-zhi[2], CHEN Liang[2], ZHENG Zi-yan[2], ZHAO Na[6]

(1. Beijing Information Science and Technology University, Beijing 100101*, China;2. The Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing* 100029*, China;3. University of Chinese Academy of Sciences, Beijing* 100049*, China;4. School of Science, Beijing University of Posts and Telecommunications, Beijing* 100876*, China;5. Peking University, Beijing* 100871*,China;6. Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing* 100101*, China)*

Abstract: In this study, the daily maximum temperature observations of 65 meteorological stations in Central Asia, ERA-Interim reanalysis data, and latitude, longitude, altitude data were used to construct a random forest interpolation model. Based on this model, we filled the missing daily maximum temperature data ($T_{\mathrm{Station_f}}$) of the meteorological stations and constructed the daily maximum temperature grid data set ($T_{\mathrm{RFIM_G}}$) for Central Asia with a spatial resolution of 0.75°×0.75 ° from 1979 to 2016. Based on $T_{\mathrm{RFIM_G}}$, the trends of extreme high temperature in Central Asia in summer from 1979 to 2016 were then analyzed. The result shows that the regional average extreme high temperature indices increase rates range from 0.22℃·(10a)$^{-1}$ to 0.30℃·(10a)$^{-1}$, and the significant warming areas are mainly distributed in western Kazakhstan, most of Turkmenistan, and southeastern Uzbekistan. The increase rates of summer extreme high temperature indices based on T_{RFIM_G} are significantly greater than those based on $T_{\mathrm{Station_f}}$. This indicates that the estimation of extreme summer high temperature in this region using the station observations is significantly underestimated. The data set ($T_{\mathrm{RFIM_G}}$) obtained in this study can to some extent compensate for the shortcomings of using the station observations to partly describe the extreme high temperature changes in Central Asia, so as to correctly guide people to take corresponding mitigation and adaptation measures in response to extreme weather and climate events.

Key words: random forest interpolation; machine learning; extreme high temperature in summer; Central Asia